

Evolutionarily conserved and functionally important residues in the I-CeuI homing endonuclease

Monique Turmel*, Christian Otis, Vincent Côté and Claude Lemieux

Program in Evolutionary Biology, Canadian Institute for Advanced Research, Département de Biochimie, Faculté des Sciences et de Génie, Université Laval, Québec, Québec G1K 7P4, Canada

Received March 12, 1997; Accepted April 29, 1997

DDBJ/EMBL/GenBank accession nos*

ABSTRACT

Two approaches were used to discern critical amino acid residues for the function of the I-CeuI homing endonuclease: sequence comparison of subfamilies of homologous proteins and genetic selection. The first approach revealed residues potentially involved in catalysis and DNA recognition. Because I-CeuI is lethal in *Escherichia coli*, enzyme variants not perturbing cell viability were readily selected from an expression library. A collection of 49 variants with single amino acid substitutions at 37 positions was assembled. Most of these positions are clustered within or around the LAGLI-DADG dodecapeptide and the TQH sequence, two motifs found in all protein subfamilies examined. The K_m and k_{cat} values of the wild-type and nine variant enzymes synthesized *in vitro* were determined. Three variants, including one showing a substitution of the glutamine residue in the TQH motif, revealed no detectable endonuclease activity; five others showed reduced activity compared to the wild-type enzyme; whereas the remaining variant cleaved the top strand about three times more efficiently than the wild-type. Our results not only confirm recent reports indicating that amino acids in the LAGLI-DADG dodecapeptide are functionally critical, but they also suggest that some residues outside this motif directly participate in catalysis.

INTRODUCTION

Homing endonucleases confer mobility to the DNA sequences that encode them by producing double-stranded breaks in their target sites (reviewed in 1,2). The coding sequences of homing endonucleases usually interrupt genes and consist either of open reading frames (ORFs) in introns or of inframe spacers in protein-coding sequences (also designated as intein coding sequences), and their target sites are the cognate intronless or spacerless alleles. Despite their low degree of sequence similarity, the majority of homing endonucleases can be divided into four major structural families based on the presence of the following

amino acid motifs: variations of LAGLI-DADG (a dodecapeptide), GIY-YIG, H-N-H and the His-Cys box (reviewed in 3). Most of the homing endonucleases that have been identified thus far belong to the family of LAGLI-DADG proteins. Included within this family are proteins associated with functions other than double-strand DNA cleavage (e.g. RNA maturation and protein processing).

Like bacterial restriction endonucleases of type I and II, homing endonucleases require Mg^{2+} as a co-factor, but they clearly differ from these bacterial enzymes in recognizing asymmetric and long DNA sequences (1,2). The target sites of homing endonucleases range from 15 bp for I-PpoI to 37 bp for I-TevI. All of these enzymes introduce a staggered cut into a double-stranded DNA, with the exception of two enzymes (I-HmuI and I-HmuII) which nick only one DNA strand (4). The LAGLI-DADG endonucleases as well as I-PpoI (His-Cys box) generate 4 nucleotide (nt) 3'-OH overhangs at or near the intron insertion site, whereas the bacteriophage T4 endonucleases I-TevI (GIY-YIG), I-TevII (unknown motif) and I-TevIII (H-N-H) leave 2 nt 3'-OH overhangs or 2 nt 5'-OH overhangs at a distance from the point of intron insertion. A number of LAGLI-DADG endonucleases (I-CeuI, 5; I-SceI, 6; I-ChuI, 7; I-SceIII, 8; and I-CpaI, 9) preferentially cleave one of the two DNA strands (top or bottom strand) of the target sequence; as demonstrated for I-TevI (10) and I-TevII (12), cleavage of one strand is most probably required for cleavage of the other strand.

The distinctive properties of the homing endonucleases raise considerable interest in understanding how these enzymes work. Our group is particularly interested in identifying the functional domains of the I-CeuI endonuclease. This protein of 218 amino acids is encoded by the fifth group I intron (CeLSU-5, subclass IB4) in the chloroplast large subunit (LSU) rRNA gene of the green alga *Chlamydomonas eugametos* (12,13). It features only one of the two dodecapeptides generally found in LAGLI-DADG proteins and it displays good sequence similarity over most its length with a small number of proteins from this family which are encoded by introns inserted within different sites of the LSU rRNA gene or within other genes (12,14). I-CeuI recognizes a degenerate sequence of 17–19 bp to produce a staggered cut 5 bp downstream from the CeLSU-5 intron insertion site (5,15). Its target DNA sequence is highly conserved, being found in the LSU

*To whom correspondence should be addressed. Tel: +1 418 656 2131; Fax: +1 418 656 7176; Email: mturmel@rsvs.ulaval.ca

+L42859, L42860, L42984, L42986, L43360, L43503, L44123, L49149

rRNA gene of a wide spectrum of organisms. I-CeuI cleaves the DNA of enteric bacteria and *Bacillus subtilis* within each of the LSU rRNA genes of the *rrn* operons (16,17). Because of this great specificity and the high conservation of the rRNA gene loci, this homing endonuclease has proved very useful for the mapping of bacterial genomes (16,18 and refs therein, 19–22).

The present study was undertaken in order to discern the amino acid residues critical for the function of I-CeuI. Comparative sequence analysis of subfamilies of intron-encoded proteins homologous to I-CeuI, including the subfamily of endonucleases that cleave the same DNA substrate as this enzyme, revealed the evolutionarily conserved residues that are potentially involved in catalysis and DNA recognition. On the other hand, a genetic approach taking advantage of the lethality of I-CeuI in *Escherichia coli* (13,15, this study) allowed us to identify substitution sites in variant enzymes with reduced cleavage activity. Our results not only confirm recent reports indicating that residues in the LAGLI-DADG dodecapeptide are functionally important, but also suggest that some residues outside this motif directly participate in catalysis.

MATERIALS AND METHODS

Identification and sequence analysis of LSU rDNA introns and their encoded proteins

The regions spanning the CeLSU-5 intron insertion site within the chloroplast LSU rRNA genes of various green algae (their source is indicated in Fig. 1) were amplified by PCR. The strains were grown in modified Volvox medium (23) or medium K (24). The PCR amplifications were performed using the primers d(ACAGGTCTCCGCAAAGTCGTA) (#53) and d(TGACCGAGTCTCTCTCCGAGAC) (#850) and total cellular DNA or chloroplast DNA (cpDNA)-enriched preparations which were obtained as outlined in refs 25 and 26, respectively. The PCR products were sequenced using the dsDNA cycle sequencing system from Life Technologies Inc./BRL (Canada). ³²P-labeled primers that are complementary to highly conserved LSU rRNA gene regions as well as intron-specific oligonucleotides were employed to initiate the sequencing reactions. Sequence analysis was mainly carried out with the Genetics Computer Group software (27). The programs PILEUP, PROFILEMAKE, PROFILESEARCH and PROFILESEGMENTS in this package were employed to search proteins in databases which show a sequence profile similar to a group of related intron-encoded proteins. The multiple protein alignments shown in Figures 2 and 4 were generated using CLUSTALW (28); analyses of these alignments were performed with the computer program AMAS (29), and the outputs were represented using ALSCRIPT (30).

Assays of endonuclease activity

Intron-encoded proteins were synthesized *in vitro* and assayed for endonuclease activity as outlined previously (14) with the following modifications. Regions encompassing the intron ORFs were amplified by PCR from total cellular or cpDNA-enriched preparations using the following primers: for the CecLSU-2, CelLSU-1, CmuLSU-1, CluLSU-1 and AstLSU-1 ORFs, d(TAATACGACTCACTATAGGGAGATAGGGCAATCAGCAGGAAAC) (#762) and d(AGGATGACGTATAGTCTCTGA) (#766); for the CeuLSU-5 ORF, d(TAATACGACTCACTATAGGGAGACCGAGTA-

GGTGACACGCTAA) (#621) and d(ATTACGCCTTTCGTGCAGGCC) (#56); for the CmoLSU-1 ORF, d(TAATACGACTCACTATAGGGAGATAGGGCAATCAGCAGGAAAT) (#622) and d(AGAGTAACGTGTAGTCTCTGA) (#788); for the CpaLSU-1 ORF, d(TAATACGACTCACTATAGGGAGATAGGGCAATCAGCAGGAAAT) (#622) and d(TGCTTGAACCTCCGTTCCGGCG) (#623); and for the SobLSU-1 ORF, d(TAATACGACTCACTATAGGGAGAAAGACCAATCAGCAGGAAAC) (#787) and d(TATAAAACGTGTAGTCTCTGA) (#786). The DNA substrate was generated by PCR from a *Chlamydomonas moewusii* (UTEX 97) cpDNA-enriched preparation (100 ng) with the ³²P-labeled primers d(ACAGGTCTCCGCAAAGTCGTA) (#53) and d(AGTCCGCATCTTCACGGGACA) (#517).

Isolation of *E.coli* clones expressing variant I-CeuI enzymes

The I-CeuI gene was PCR-amplified from a *C.eugametos* (UTEX 9) cpDNA-enriched preparation (100 ng) using the primers d(CCAAATAACCCATGGCAAACCTTTA) (#149) and d(GTCTCTGAGGATCCTACTTTATAC) (#835), and cloned into the T7 promoter expression plasmid pET-30a(+) (Novagen). These primers introduced *Nco*I and *Bam*HI restriction sites (underlined) at the ends of the I-CeuI gene coding sequence in order to clone the latter sequence into the respective sites of pET-30a(+), thus fusing the I-CeuI gene in-frame with the sequences coding for the His-tag and S-tag on the plasmid. This strategy resulted in a change of codon at the second position (S→A) in the I-CeuI gene. The amplification reaction was carried out for 25 cycles (1 min at 94°C, 1 min at 50°C, and 2 min at 72°C) followed by a 25 min incubation at 72°C in 100 µl of buffer consisting of 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 200 mM of each dNTP, 2.5 mM MgCl₂, and 2.5 U of AmpliTaq DNA polymerase (Perkin-Elmer). The *E.coli* strain NovaBlue (DE3) (Novagen) was transformed with the ligation mixtures and plated on LB medium containing 30 µg/ml kanamycin. Plasmid DNA was isolated from individual transformants, and the presence of inserts was tested by restriction analysis with *Nco*I and *Bam*HI. Inserts were sequenced using the dsDNA cycle sequencing system of Life Technologies Inc./BRL (Canada) and the following primers: d(TAGCACTAAAAAGCTCGCTAC) (#113), d(TGCCTTCA-GTTCTCCAG) (#156) and d(AACCCCATTTGACATGTTGA-GTC) (#998).

Determination of the kinetic parameters of variant I-CeuI enzymes

To test the activity of variant I-CeuI enzymes, these were synthesized *in vitro* as follows. Prior to the *in vitro* transcription and translation reactions, the I-CeuI gene coding region in each recombinant pET-30a expression plasmid was transferred as a *Nco*I–*Bam*HI fragment into the pCITE-3a(+) plasmid (Novagen). This T7 promoter-containing plasmid features a CAP-independent translation enhancer that allows the optimal production of proteins *in vitro*. The segment of each resulting recombinant plasmid, which is comprised between the position –43 relative to the T7 promoter and the position +150 relative to the stop codon of the I-CeuI gene coding region, was PCR-amplified from the ligation mixture using the primers d(GTTTTCCAGTCACGACGTTGT) (#956) and d(CCTTTCGGGCTTTGTTAGCAG) (#1001). After purification, the amplified DNA fragment (250 ng) was transcribed and the resulting RNA was translated using the Single Tube Protein System 2 of Novagen. The wild-type

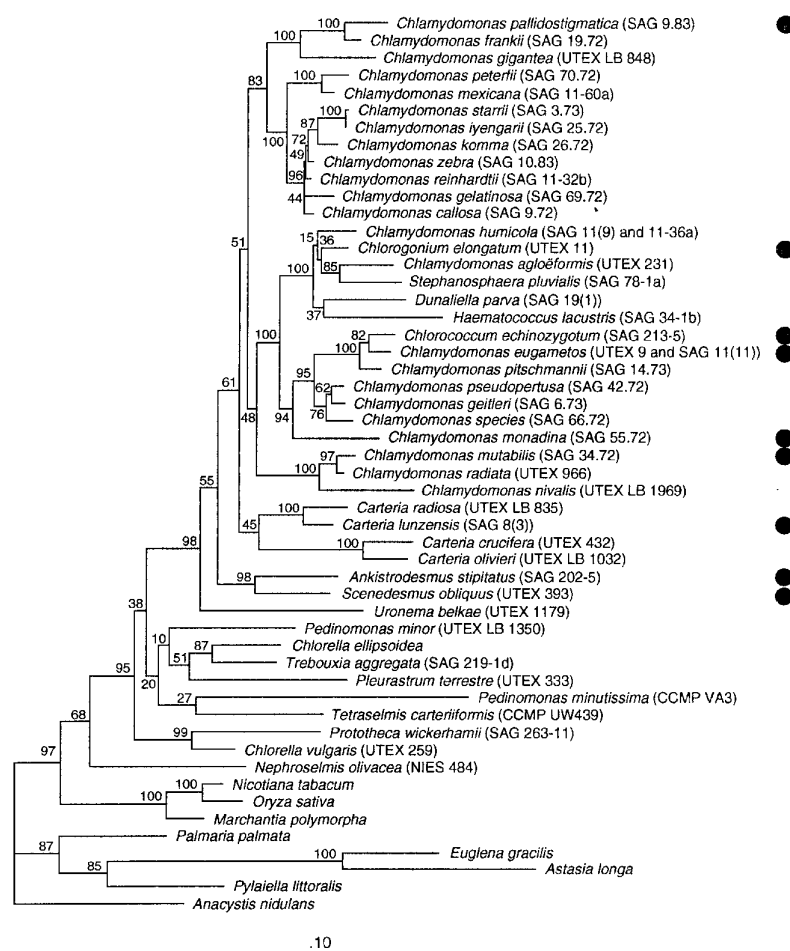


Figure 1. Distribution of chloroplast group I introns at site 1923 of the LSU rRNA gene in a variety of photosynthetic eukaryotes. The taxa featuring introns are denoted by circles. The phylogeny was inferred from chloroplast LSU rRNA gene sequences with fastDNAm1 (31), using the *Anacystis nidulans* LSU rRNA gene sequence as an outgroup (14). The percentage of bootstrap resamplings (100 replicates) that support the topological elements is shown to the left of each grouping. The scale bar indicates 10 changes/100 nt. The LSU rRNA genes of the taxa whose origins are indicated in parentheses were sequenced in our laboratories as described in 'Materials and Methods'. CCMP, Center for the Culture of Marine Phytoflagellates at the Bigelow Laboratory for Ocean Sciences; NIES, National Institute for Environmental Studies at Tsukuba; SAG, Sammlung von Algenkulturen at Göttingen; UTEX, Culture Collection of Algae at the University of Texas.

I-CeuI endonuclease was also synthesized *in vitro* as aforementioned, except that the recombinant pCITE3a(+) construct was obtained by ligating a PCR product that was generated from a *C.eugametos* cpDNA-enriched preparation with the primers #149 and #835. The relative concentrations of *in vitro*-synthesized proteins were determined using the S-Tag Assay Kit of Novagen. This is a spectrophotometric assay based on the interaction of the 15 amino acid S-Tag peptide with the ribonuclease S-protein; these two components form an active ribonuclease whose activity is measured using poly(C) as substrate. The relative concentrations of S-Tag fusion I-CeuI proteins were confirmed by measuring the incorporation of [³⁵S]methionine.

The endonuclease activity of each *in vitro*-synthesized protein was assayed at 37°C in a 40 µl reaction mixture containing 20 mM Tris-HCl (pH 8.5), 2.5 mM MgCl₂, 1 mM dithiothreitol and various amounts (0.25–100 fmol/µl) of a ³³P-labeled, PCR-amplified fragment that was generated from a *C.moewusii* cpDNA-enriched preparation with the primers #53 and #517. For this PCR reaction, 50 ng of each of the primers were 5'

end-labeled using [γ-³³P]ATP (25 mCi, 6000 Ci/mmol) and T4 polynucleotide kinase, and these labeled primers were added to the reaction mixture along with 350 ng of each of the unlabeled primers. Cleavage reactions were initiated by the addition of 4 µl of the translation mixture. Following incubation periods of 10 and 20 min, a 20 µl aliquot was removed from each reaction mixture and added to a tube containing 1 µl of 0.5 M EDTA and 1.25 µl of 10% SDS. All reaction samples were incubated at 50°C for 5 min, then 1.75 µl H₂O and 1 µl of 5 µg/µl proteinase K were added, and incubation was continued at 50°C for 1 h. After ethanol precipitation in the presence of ammonium acetate and glycogen (20 µg), DNA was dissolved in 5 µl sterile H₂O, and then 2.5 µl of loading buffer [95% (v/v) formamide, 20 mM Na₂EDTA, 0.05% (w/v) bromophenol blue, 0.02% (w/v) xylene cyanol FF] was added. Approximately equal amounts of all DNA samples were electrophoresed in 5% acrylamide–9 M urea gels. The fraction of total DNA represented by each DNA product was quantified with the Fujix BAS 1000 Bio Imaging Analyzer (Fuji) using the MacBAS image analysis software.

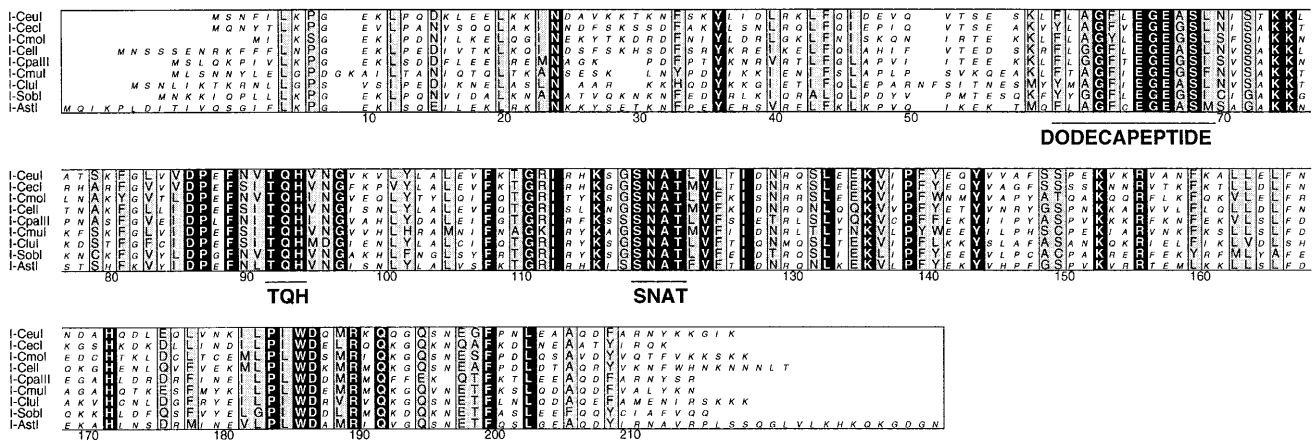


Figure 2. Sequence alignment of the predicted proteins encoded by chloroplast group I introns at site 1923 of the LSU rRNA gene. Identical residues are shown on a black background, while conserved sets of amino acids sharing eight of the 10 features in the property matrix intra.pt of the AMAS program (29) are shown on a grey background. Horizontal lines below the alignment denote the regions containing the conserved LAGLI-DADG dodecapeptide and the TQH and SNAT motifs. The proteins all possess endonuclease activity (see Fig. 3) and are thus designated according to the nomenclature proposed by Dujon *et al.* (32). Database accession numbers for the sequences are: I-CeuI, Z17234; I-CecI, L44123; I-CmoI, L49149; I-Cell, L42860; I-CpaIII, L43503; I-CmuI, L42859; I-ChuI, L42986; I-SobI, L43360; and I-AstI, L42984.

RESULTS

Amino acid residues conserved among homing endonucleases that cleave the same DNA substrate as I-CeuI

Sequence comparison of subfamilies of homologous proteins which differ in substrate specificity is a powerful method to predict the amino acid residues involved in biological function and substrate specificity. In the case of the I-CeuI endonuclease, use of this approach required the identification of homologues that cleave the same substrate as this enzyme. As such homologues were likely to be encoded by LSU rDNA introns found at the same position as the CeLSU-5 intron, we searched for coding sequences in introns located at this position, i.e., at position 1923 relative to the *E. coli* 23S rRNA. Our search was carried out in the course of a more global study of the chloroplast LSU rRNA gene, which aimed to elucidate the phylogenetic relationships among green algae by comparative analysis of the rRNA-encoding sequences, and also to probe the evolution of group I introns in this gene. The complete sequences of the chloroplast LSU rRNA genes of 48 green algae are available at present. All of these sequences except one (that of *Chlorella ellipsoidea*) were determined in our laboratories (M.T. and C.L.). Only the sequences of the rRNA-encoding regions and of some introns from 17 *Chlamydomonas* species have been described thus far in the literature (7,12,14,26). We report below the distribution and features of the proteins encoded by the introns at site 1923.

As shown in Figure 1, aside from *C.eugametos*, eight green algae from the class Chlorophyceae were found to harbour a group I intron at site 1923 of the chloroplast LSU rRNA gene (Fig. 1). All of the eight newly identified introns share not only a core structure that is remarkably similar to that of CeLSU-5 (data not shown), but also an ORF of more than 210 codons with a sequence similar to the CeLSU-5 ORF (see Fig. 2). As in the CeLSU-5 intron, the ORFs of all the introns reside in the loop extending P6. An alignment of their predicted amino acid sequences with that of the I-CeuI endonuclease is shown in Figure 2. In pairwise sequence comparisons, the level of identity (and similarity) displayed by the nine site-1923 intron-encoded

proteins ranges from 39.5% (59%) to 62% (76.5%). All of the nine proteins can be aligned over their entire length, except for the amino acid residues at the N- and C-termini, and only four gaps of one to three amino acids are observed in the alignment. Forty positions show strictly conserved residues and 80 show similar amino acids. The highest densities of strictly conserved residues are found in the LAGLI-DADG dodecapeptide and two other regions designated the TQH and the SNAT motifs (see Fig. 2). Each of these three regions displays at least three consecutive strictly conserved positions. The dodecapeptide motif is the most conserved region, with five positions that are absolutely conserved and three that show similar amino acids.

To determine whether each of the eight newly identified site-1923 introns encodes a double-strand DNA endonuclease that is specific for the DNA substrate of I-CeuI, we undertook the synthesis of these intron-encoded proteins *in vitro* and tested their endonuclease activity in cleavage assays. Proteins of appropriate sizes were produced *in vitro* by T7 RNA polymerase-mediated transcription of PCR-amplified fragments carrying the various intron ORFs and subsequent translation of the resulting RNAs in rabbit reticulocyte lysates. Each of the translation products was tested in a cleavage assay with a ³²P-labeled 258 bp PCR product derived from a segment of the *C.moewusii* chloroplast LSU rRNA gene which corresponds to positions 1789–2041 in the *E. coli* 23S rRNA. This intronless DNA segment shows perfect sequence identity with the exon sequences immediately adjacent to the site-1923 introns in the chloroplast LSU rRNA genes of *C.eugametos*, *Chlorococcum echinozygotum*, *Chlorogonium elongatum* and *Chlamydomonas mutabilis*. The equivalent sequence in the vicinity of each of the five other site-1923 introns displays one to three substitutions (data not shown).

The proteins encoded by all eight newly identified intron ORFs, except that originating from *Chlamydomonas monadina* (or I-CmoI), were found to cleave specifically the DNA substrate at the same position as I-CeuI (i.e. 5 bp downstream of the site at which the intron is inserted in the intron-plus allele), yielding fragments of 144 and 118 bp (Fig. 3). The 144 bp fragment corresponding to the top strand was more abundant than the 118 bp fragment corresponding to the other strand (Fig. 3), indicating

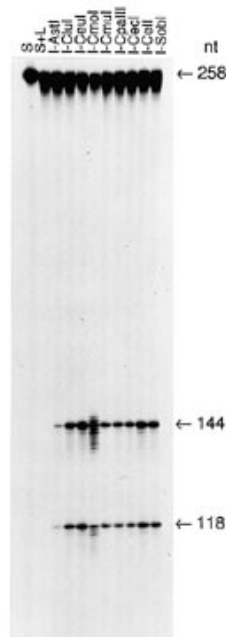


Figure 3. Enzymatic activities of the endonucleases encoded by chloroplast group I introns inserted at site 1923 of the LSU rRNA gene. An unprogrammed rabbit reticulocyte lysate (S+L) and rabbit reticulocytes containing *in vitro*-synthesized RNAs from the intron ORFs were incubated at 37°C for 1.5 h in 50 μ l of a reaction mixture containing 20 mM Tris-HCl (pH 8.0), 2.5 mM MgCl₂, 1 mM dithiothreitol, and a ³²P-labeled intronless DNA substrate (258 bp) derived from the *C.moewusii* chloroplast LSU rRNA gene. The reaction products were electrophoresed in 5% polyacrylamide/urea gels alongside the corresponding DNA substrate (S).

that, as previously reported for I-CeuI (5), the top strand is preferentially cleaved. The much weaker intensities of the cleavage products generated by the *Ankistrodesmus stipitatus* endonuclease (I-AstI) are undoubtedly due to the low yield of protein produced during the *in vitro* translation reaction; the incorporation of [³⁵S]methionine measured during this reaction revealed a reduction of ~10-fold in the amount of the I-AstI protein compared to the other intron-encoded proteins. The cleavage assay with the *C.monadina* protein (I-CmoI) yielded a few fragments in addition to the expected 144 and 118 bp products (Fig. 3). While some of these fragments could be the result of 3' exonuclease activity, the 145 and 146 bp fragments cannot be degradation products and must originate from the cleavage of the top strand at 6 and 7 bp downstream from the intron insertion site. Cleavage assays using two different concentrations of MgCl₂ (2.5 and 5 mM) in the presence or absence of NaCl (25 mM) did not alter the cleavage pattern.

Sequence comparison of the I-CeuI subfamily of homing endonucleases with related subfamilies of LAGLI-DADG proteins

To discriminate between amino acid residues that may be involved in catalysis and those that may be involved in recognition of the DNA substrate, we have compared the sequences of I-CeuI and of its homologues encoded by site-1923 introns with those of related LAGLI-DADG proteins showing similarity over a long segment outside of the dodecapeptide motif (see Fig. 4). The latter proteins were identified by searching the

SwissProt and PIR databases using a profile of the subfamily of site-1923 intron-encoded proteins. This analysis revealed several proteins that were not included in a previously published alignment containing the I-CeuI sequence (12); our greatest surprise was to find two proteins encoded by group I LSU rDNA introns inserted at site 2593, the I-CreI endonuclease from the *Chlamydomonas reinhardtii* chloroplast (33) and its homologue in *Acanthamoeba castellanii* mitochondria (34), as well as the protein encoded by the *A.castellanii* group I intron inserted at site 1951 in the mitochondrial LSU rRNA gene (34). The updated, multiple alignment shown in Figure 4 contains four subfamilies of proteins encoded by LSU rDNA introns and a group of proteins encoded by introns found in various mitochondrial genes. Aside from I-CreI, I-CeuI and the eight endonucleases identified in the course of this study, I-CpaI from the *Chlamydomonas pallidostigmatica* chloroplast is the only protein in the alignment which has been shown to possess endonuclease activity (14).

The endonuclease residues playing a role in catalysis and Mg²⁺ binding are predicted to be completely or almost completely conserved among the intron-encoded proteins compared in Figure 4. In addition to the residues G⁶¹, F⁶², G⁶⁵ and E⁶⁶ in the LAGLI-DADG motif of I-CeuI, the residues F⁸⁹, Q⁹³ and G¹¹¹ are candidates for such a role. On the other hand, the residues participating in the recognition and binding of the DNA substrate in a given endonuclease are predicted to be characteristic of the protein subfamily to which this endonuclease belongs. Such residues are relatively numerous (~20) in I-CeuI and tend to be clustered in the second half of the protein (Fig. 4). Of particular interest is the motif SNAT which corresponds to a gap in most proteins from other subfamilies.

Analysis of mutations among I-CeuI gene sequences expressed in *E.coli*

We have previously encountered major problems in recovering stable clones of *E.coli* producing large amounts of the I-CeuI endonuclease from expression plasmids. This endonuclease proved highly toxic to *E.coli* when we expressed the I-CeuI gene using the *trc* promoter expression vectors pKK233-2 and pTRC-99A (13) as well as other expression plasmids (15): high frequency of frameshift mutants were observed, extremely low levels of expression were obtained, and growth of *E.coli* was found to be greatly affected. As the I-CeuI endonuclease has been shown to cleave the *E.coli* chromosomal DNA specifically within the rRNA operons (16), the toxicity of I-CeuI is very likely to be a consequence of such cleavage. Assuming that this hypothesis is valid, it should be possible to easily isolate mutants in the I-CeuI gene by searching for clones whose growth is not impaired by the expression of this gene.

With the goal of isolating such mutants, we undertook the construction of a library of *E.coli* clones expressing the I-CeuI gene. This gene was amplified by PCR using conditions that are expected to yield a frequency of mutations of 0.25–0.4% (36) and it was cloned into the *Nco*I and *Bam*HI sites of the T7 expression plasmid pET-30a(+) in order to fuse the I-CeuI gene in-frame with the sequences coding for the His- and S-tags. The recombinant plasmids were then introduced into the *E.coli* strain NovaBlue (DE3) containing a chromosomal copy of the T7 RNA polymerase gene under the control of the *lacUV5* promoter and transformants were selected in the absence of IPTG. Even under these conditions in which transcription from both the *lacUV5* and T7 promoters is inhibited, the I-CeuI endonuclease was produced in amounts that are

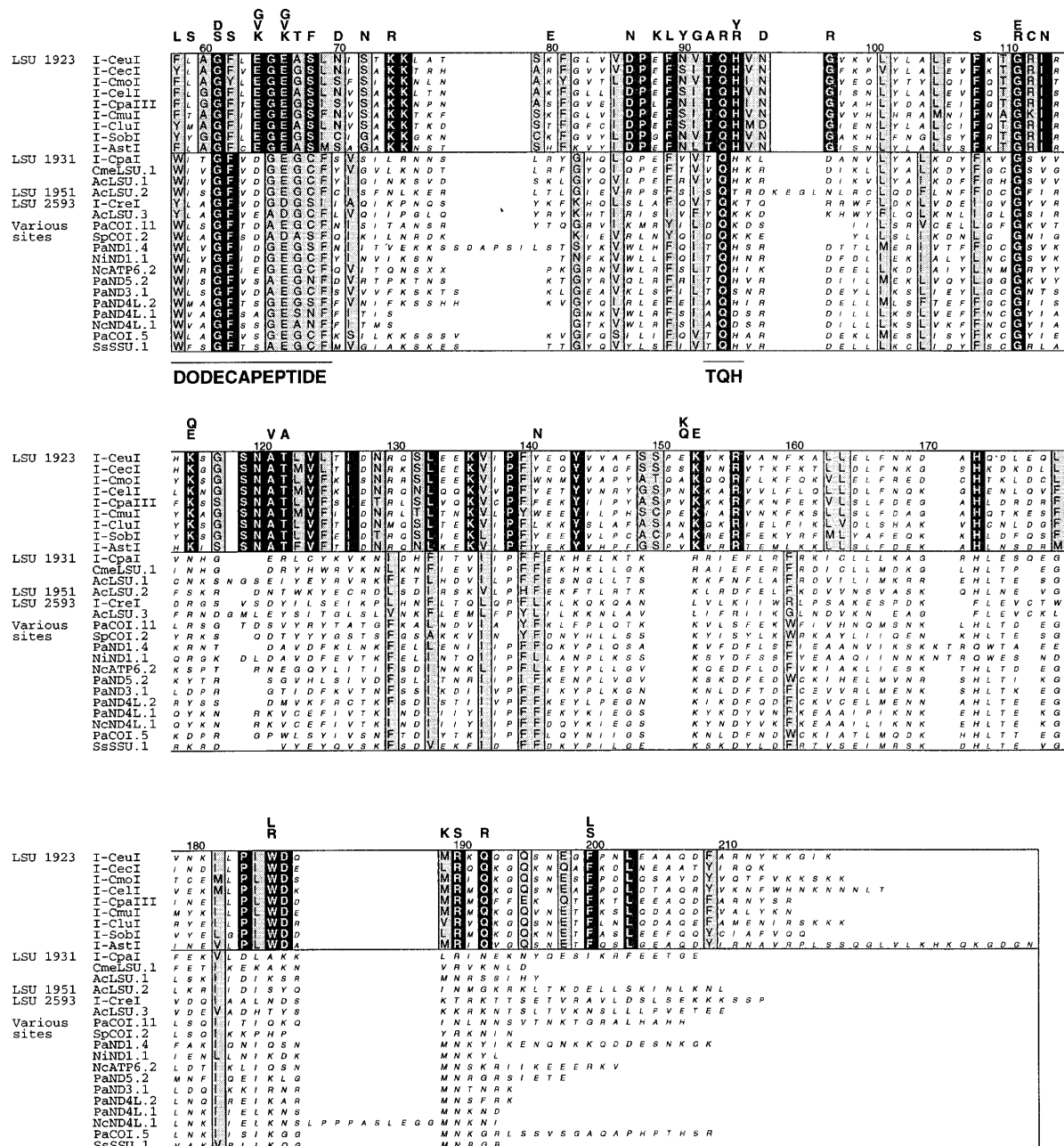


Figure 4. Sequence alignment of all known group I intron-encoded proteins that are related to I-CeuI. The compared proteins are encoded by introns inserted at sites 1923, 1931, 1951 and 2593 in the LSU rRNA gene and also by mitochondrial introns found in other genes. Identical residues are shown on a black background, while conserved sets of amino acids sharing eight of the 10 features in the property matrix intra.pt of the AMAS program (29) are shown on a grey background. Note that in the calculation of conservation values up to 8% of atypical residues at any position of each group of sequences were ignored. The coordinates refer to the I-CeuI protein. The letters above the alignment indicate the single amino acid substitutions that are described in Table 1. The proteins known to display endonuclease activity are designated according to Dujon *et al.* (32); the introns encoding the other proteins are abbreviated according to Michel and Westhof (35), except that OXI was replaced by COI and that a three-letter code, instead of a two-letter code, was used to designate *Chlamydomonas mexicana*. All protein sequences are partial; in each sequence, the segment preceding the single or the second copy of the dodecapeptide was deleted. All of the LSU rRNA intron-encoded proteins as well as the PaND4-1-encoded protein contain a single copy of the dodecapeptide, whereas the remaining ones contain two copies of this sequence. Database accession numbers for the sequences: I-CpaI, L36830; CmeLSU-1, L49148; AcLSU-1, s53825; I-CreI, a23091; AcLSU-3, s53827; NcATP6-2, X01075; PaCOI-5, f48327; SpCOI-2, P22190; NiND1-1, s06367; PaND1-4, s06059; PaND3-1, s05654; PaND4-1, s05653; NcND4L-1, s10840; PaND4L-1, s09134; PaND4L-2, S09141; NcND5-1, s10841; PaND5-2, s09143; SsSSU-1, U07553; the remainder are given in the legend of Figure 1.

lethal for *E. coli* cells. The efficiency of transformation was several orders of magnitude lower than expected, and all 133 recombinant clones analyzed were found to display mutations in their I-CeuI gene sequence. Because we also failed to recover the wild-type form of

the I-CeuI gene by transforming a NovaBlue *E. coli* strain lacking the T7 RNA polymerase gene with our plasmid construct, it is likely that transcription of the I-CeuI gene occurred from cryptic promoters present on the pET30(a+) plasmid.

Table 1. Nature and sites of the mutations responsible for the single amino acid changes in I-*CeuI* variants

Codon	Codon position	Base change	Type of mutation ^a	Amino acid change	Independent clones
58	1	TTT→ <u>C</u> TT	Ts	F→L	1
59	2	TTA→TCA	Ts	L→S	1
61	1	GGT→ <u>A</u> GT	Ts	G→S	1
61	2	GGT→G <u>A</u> T	Ts	G→D	1
62	2	TTT→T <u>C</u> T	Ts	F→S	4
64	1	GAA→ <u>A</u> AA	Ts	E→K	3
64	2	GAA→G <u>G</u> A	Ts	E→G	1
64	2	GAA→G <u>C</u> A	Tv	E→G	1
64	2	GAA→G <u>T</u> A	Tv	E→V	1
66	1	GAA→ <u>A</u> AA	Ts	E→K	1
66	2	GAA→G <u>T</u> A	Tv	E→V	2
66	2	GAA→G <u>G</u> A	Ts	E→G	1
67	1	GCT→ <u>A</u> CT	Ts	A→T	1
68	2	TCT→T <u>T</u> T	Ts	S→F	1
70	1	AAT→ <u>G</u> AT	Ts	N→D	1
72	2	AGC→A <u>A</u> C	Ts	S→N	1
74	2	AAA→A <u>G</u> A	Ts	K→R	1
80	1	AAA→ <u>G</u> AA	Ts	K→E	1
86	1	GAT→ <u>A</u> AT	Ts	D→N	1
88	1	GAA→ <u>A</u> AA	Ts	E→K	1
89	1	TTC→ <u>C</u> TC	Ts	F→L	1
90	1	AAT→ <u>T</u> AT	Tv	N→Y	1
91	2	GTG→G <u>G</u> G	Tv	V→G	1
92	1	ACT→ <u>G</u> CT	Ts	T→A	1
93	2	CAA→C <u>G</u> A	Ts	Q→R	3
94	1	CAT→ <u>T</u> AT	Ts	H→Y	1
94	2	CAT→C <u>G</u> T	Ts	H→R	2
96	1	AAT→ <u>G</u> AT	Ts	N→D	1
97	1	GGG→ <u>A</u> GG	Ts	G→R	1
108	2	TTT→T <u>C</u> T	Ts	F→S	1
111	1	GGG→ <u>A</u> GG	Ts	G→R	1
111	2	GGG→G <u>A</u> G	Ts	G→E	1
112	1	CGT→ <u>T</u> GT	Ts	R→C	1
113	2	ATT→A <u>A</u> T	Tv	I→N	1
116	1	AAA→ <u>G</u> AA	Ts	K→E	1
116	1	AAA→ <u>C</u> AA	Tv	K→Q	1
121	2	GCA→G <u>T</u> A	Ts	A→V	1
122	1	ACT→ <u>G</u> CT	Ts	T→A	1
141	1	TAT→ <u>A</u> AT	Tv	Y→N	1
152	1	GAA→ <u>C</u> AA	Tv	E→Q	1
152	1	GAA→ <u>A</u> AA	Ts	E→K	1
153	1	AAA→ <u>G</u> AA	Ts	K→E	1
186	1	TGG→ <u>C</u> GG	Ts	W→R	4
186	2	TGG→T <u>T</u> G	Tv	W→L	1
189	2	ATG→A <u>A</u> G	Tv	M→K	1
190	1	CGT→ <u>T</u> GT	Ts	R→S	1
192	2	CAA→C <u>G</u> A	Ts	Q→R	3
200	2	TTT→T <u>C</u> T	Ts	F→S	1
200	3	TTT→TT <u>A</u>	Tv	F→L	1

^aTs, transition; Tv, transversion.

Sixty eight percent of the recombinant clones (90/133) revealed single- or double-point mutations in the *I-CeuI* gene, whereas the others displayed frameshift mutations. A total of 66 clones (or 50%) harboured single-point mutations: 58 of these mutations are responsible for amino acid changes at 35 positions, while the remaining ones account for the creation or deletion of stop codons at distinct sites. Considering the five double mutants exhibiting one synonymous and one non-synonymous mutation, a total of 49 variant enzymes showing distinct, single amino acid changes at 37 positions were recovered (Table 1 and Fig. 4). All but four of these positions (positions 59, 80, 88 and 152) coincide with sites that are evolutionarily conserved within the *I-CeuI* protein subfamily (see Figs 3 and 4). Most of the amino acid changes (37/49) are found in the portion of the protein occupied by the first 123 amino acids and they are clustered in the regions containing the dodecapeptide, TQH motif and G¹¹¹. At the gene level, the great majority of the substitutions occurred at first and second codon positions, and transitions were found to be more frequent (3.1 times more numerous) than transversions (Table 1). Assuming that all mutations were caused by transitions, two amino acid substitutions would be typically predicted at each altered position. Saturation in the number of transition events was observed at only five of the 37 positions that revealed amino acid changes. To reduce the bias in the ratio of transitions and transversions, we tried to amplify the *I-CeuI* gene in the presence of 0.5 mM MnCl₂; however, these conditions which are known to yield an equal ratio of transitions and transversions and reduce the fidelity of DNA synthesis (36) did not allow us to diversify our collection of mutants with single substitutions since the frequency of mutants with multiple mutations rose sharply.

Kinetic properties of *I-CeuI* variant endonucleases

Each of the *I-CeuI* variant enzymes is predicted to be altered in its kinetic properties relative to the wild-type form. To verify this prediction and hence confirm that our *E.coli* expression library can provide valuable information on the structure/function relationships of *I-CeuI*, we have determined the k_{cat} and K_{m} values of nine randomly selected variants showing single amino acid changes at distinct positions (Table 2). These enzymes as well as the wild-type form were synthesized *in vitro* in rabbit reticulocytes lysates and assayed at 37°C in the presence of 2.5 mM MgCl₂ and various concentrations of a ³²P-labeled DNA substrate. This substrate is the 258 bp PCR product that was used to determine whether the site-1923 intron-encoded proteins are endowed with endonuclease activity. The reaction products were identified and quantified using a phosphorimager. The K_{m} and k_{cat} of the wild-type enzyme were found to be 0.9 ± 0.3 nM and 3.7×10^{-5} s⁻¹, respectively. This extremely low k_{cat} value indicates that *I-CeuI* resembles the endonucleases *I-SceI* and *PI-SceI* (an intein-derived enzyme) in turning over very slowly (6,37). Table 2 shows that all nine variants, except T122A, display $k_{\text{cat}}/K_{\text{m}}$ values that are at least 10 times lower than that of the wild-type. Although the T122A mutant does not differ from the wild-type in the kinetic parameters measured, its cleavage pattern is distinct. It shows a greater preference for the top strand, with the relative amount of the reaction product corresponding to this strand being three times more abundant than that observed for the wild-type enzyme (data not shown). Three variants (E66K, Q93R and K116Q) failed to reveal any detectable cleavage activity even in the presence of a large excess of substrate.

Variants H94R and W186R feature the lowest k_{cat} value, while variant F62S, with a K_{m} of 25 nM, deviates the most from the wild-type regarding this parameter.

Table 2. Relative kinetic parameters of nine *I-CeuI* variant enzymes synthesized *in vitro*

<i>I-CeuI</i> enzyme	k_{cat}	K_{m}	$k_{\text{cat}}/K_{\text{m}}$
Wild-type ^a	1	1	1
F62S	0.5	28	0.02
E64V	0.5	7	0.07
E66K	<0.01	ND	<0.01
Q93R	<0.01	ND	<0.01
H94R	0.1	2	0.05
K116Q	<0.01	ND	<0.01
T122A	1	1	1
W186R	0.2	2	0.10
F200L	0.4	6	0.07

^aThe *I-CeuI* wild-type values for the K_{m} and k_{cat} are 0.9 ± 0.3 nM and 3.7×10^{-5} s⁻¹, respectively.

DISCUSSION

Using comparative sequence analysis and genetic selection, we have discerned a number of amino acid residues that are potentially important for the catalytic activity and DNA sequence specificity of the *I-CeuI* family of intron-encoded endonucleases. Genetic selection allowed us to constitute a collection of 49 *I-CeuI* variants at 37 distinct positions, the majority of which correspond to evolutionarily conserved sites. Our preliminary characterization of nine of the variants indicates that all of them behave differently from the wild-type enzyme with respect to the kinetics of the cleavage reaction. Three variants, each differing from the wild-type form by its charge, show undetectable endonuclease activity; five others feature reduced activity relative to the wild-type; whereas the remaining variant, located in the SNAT motif specific to the *I-CeuI* subfamily, differs from the wild-type in cleaving the top strand about three times more efficiently than the bottom strand. Taken together, our data provide a solid framework for future investigations on the structure/function relationships of *I-CeuI* and of closely related homing endonucleases.

At this time, we cannot describe precisely the specific residues that are directly involved in the function of *I-CeuI* (catalysis, recognition of DNA substrate, binding site for Mg²⁺) because several of the amino acid changes observed in the variants may have resulted in significant modification of the protein conformation, thus leading to changes in the enzyme activity. To demonstrate without any ambiguity which residues are functionally critical, analyses of enzyme kinetics and protein/DNA interaction using purified *I-CeuI* variants will be required in conjunction with site-directed mutagenesis experiments. Although such results are not available, a recent study on the yeast *PI-SceI* endonuclease by Gimble and Stephens (38) strongly suggests that the lack of *I-CeuI* activity in our E66K variant is due to the catalytic requirement of an acidic residue at the ninth position of the LAGLI-DADG dodecapeptide. Gimble and

Stephens (38) have substituted the glutamic acid residue at this position in each of the two dodecapeptide motifs of PI-*SceI* with asparagine and alanine and found that both acidic residues are essential for catalysis. In future studies with I-*CeuI*, it will be interesting to test whether the evolutionarily conserved glutamine residue in the TQH motif participates in catalysis. No cleavage activity was detected in the variant showing a substitution with arginine at this highly conserved position among the various subfamilies of homologous proteins examined. The third I-*CeuI* variant we found to display no detectable cleavage activity (K116Q) exhibits a mutation in a residue specific to the subfamily of site-1923 endonucleases; we suspect that this residue is critical for substrate recognition.

The homing endonucleases containing a single copy of the LAGLI-DADG dodecapeptide may differ from PI-*SceI* in the structure of the active site. The results of Gimble and Stephens (38) support the idea that PI-*SceI* uses a single catalytic center to cleave both DNA strands of the substrate and that the conserved aspartic acid at the ninth position in each dodecapeptide is part of the catalytic center. The role of these two acidic residues is thought to be similar to the function of the conserved acidic residues that are found in the four restriction endonucleases whose structures are known (*EcoRI*, *EcoRV*, *BamHI* and *PvuII*). In these proteins, the acidic residues as well as a conserved lysine show almost identical positions with respect to the scissile phosphodiester bond (reviewed in 39). One model predicts that the acidic residues bind to the Mg^{2+} at the active site which is believed to help stabilize the negative charges on the pentavalent transition state following nucleophilic attack by an activated water molecule. A two-metal mechanism has also been proposed for *EcoRV* and *BamHI* in which each conserved acidic residue coordinates a Mg^{2+} ion (39,40). The present data provide no clues as to how I-*CeuI* and the homologous endonucleases containing a single dodecapeptide effect cleavage of their substrate. The alignment of I-*CeuI* and its homologues reveal no evolutionarily conserved acidic amino acid outside of the dodecapeptide which could replace the acidic amino acid in the second dodecapeptide. Assuming that two acidic amino acids are required to form the active site, then the second acidic residue might not be positionally conserved among the different protein subfamilies or each endonuclease might bind to the substrate as a dimer, allowing the formation of an active site comprising an aspartic acid from each dodecapeptide. Given that I-*CeuI* has been shown to be a monomer in solution (41), the latter possibility would imply the initial formation of a protein/DNA complex comprising one bound molecule of protein followed by the binding of a second molecule of protein. Alternatively, the active site of endonucleases possessing a single dodecapeptide may show little similarity with that of restriction enzymes and contains only one acidic amino acid. The conserved glutamine residue at position 93 whose replacement with an arginine abolishes I-*CeuI* endonuclease activity could be part of such a site.

Because a recent study indicated that only one of the two dodecapeptides in I-*SceII* (a dimer) is important for endonuclease function, it has been suggested that there might be fundamental differences between intron-encoded endonucleases and their homologues derived from inteins (42). The authors of this study found that substitutions of the conserved glycines at the fourth and tenth positions of the first dodecapeptide to aspartic acid block cleavage activity, whereas the equivalent substitutions in the second dodecapeptide inhibit maturase function. Drawing an

analogy between I-*CreI* and I-*SceII*, the same authors noted that the dodecapeptide of I-*CreI* is likely to correspond to the first dodecapeptide of I-*SceII* and that this enzyme has no maturase activity, a function associated with the second dodecapeptide in I-*SceII*. The multiple protein alignment shown in Figure 4 strongly suggests that the second dodecapeptide of some LAGLI-DADG intron-encoded endonucleases could be essential for endonuclease function. All of the proteins containing a single dodecapeptide (including I-*CreI*, I-*CeuI* and I-*CpaI*) can be unambiguously aligned with the second half of their counterparts containing two copies of the dodecapeptide.

Several of the evolutionarily conserved positions in Figure 4, in particular the positions specific to the I-*CeuI* subfamily, are not represented in our collection of I-*CeuI* variants. Because the sites of amino acid substitutions are clearly not randomly distributed and also because two or more amino acid substitutions are observed at nine sites of the protein, we believe that our failure to identify amino acid changes at numerous specific positions cannot be entirely attributed to an insufficient number of the mutant clones selected for analysis of the I-*CeuI* gene sequence. It is possible that single substitutions of a number of residues involved in DNA recognition are not sufficient to yield an enzyme without any toxic effect on *E.coli*. Some of the amino acid residues that are specific to the I-*CeuI* subfamily may also be important for a function other than endonucleolytic which is not selected against in *E.coli*. One such possibility is that of a maturase function, particularly given the inability of the CeLSU-5 intron to splice *in vitro* (43), the remarkable stability of the endonuclease gene within the site-1923 LSU rDNA introns, and the larger number of conserved residues in the protein subfamily containing I-*CeuI* compared to those containing I-*CpaI* and I-*CreI*.

The *E.coli* cells carrying the five characterized I-*CeuI* variants with reduced cleavage activity can be propagated probably because the endonuclease activity is reduced to a level such that the target sites on the rDNA operons cannot be efficiently cleaved. We believe that the variant having the same K_m and k_{cat} values as the wild-type enzyme (T122A) is not lethal to *E.coli* because it more frequently cleaves the top strand. The single-strand breaks preferentially introduced by this variant can be religated readily by the *E.coli* repair system, whereas the double-strand breaks predominantly generated by the wild-type enzyme cannot.

Our inability to express the wild-type I-*CeuI* enzyme in *E.coli* contrasts with a previous report indicating that this enzyme was produced in *E.coli* (41). Another discrepancy between the results of this study and ours concerns the K_m of the wild-type enzyme. The value reported here for the *in vitro*-synthesized enzyme is 100-fold lower than the value reported for the recombinant enzyme produced in *E.coli*. In an attempt to resolve these discrepancies, we have sequenced the insertion of the expression plasmid which served to produce the recombinant enzyme and found a single mutation that causes the replacement of the arginine residue at position 156 (a residue specific to the I-*CeuI* subfamily) with a histidine. This mutation is probably responsible for the aforementioned discrepancies although we cannot eliminate the possibility that part of the reduced activity is due to the different kinds of substrate used. In addition, the presence of the His- and S-tags on the wild-type and variant enzymes studied here may alter the enzymatic properties of the endonuclease; however, such an effect proved to be negligible in recent assays with a

purified variant enzyme (N.Drouin, L.D.Eltis and C.Lemieux, unpublished results). Production of the wild-type I-CeuI enzyme in *E.coli* would require a plasmid vector completely devoid of cryptic promoters or the expression of an inactive fusion enzyme which can be cleaved with a protease to recover the active form.

ACKNOWLEDGEMENTS

We thank Lindsay D.Eltis for critical reading of the manuscript. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (to M.T. and C.L.) and from 'Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche' (to M.T. and C.L.). M.T. and C.L. are Scholars in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

- Lambowitz, A.M. and Belfort, M. (1993) *Annu. Rev. Biochem.*, **62**, 587–622.
- Mueller, J.E., Bryk, M., Loizos, N. and Belfort, M. (1993) In Linn, S.M., Lloyd, R.S. and Roberts, R.J. (eds), *The Nucleases*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY, pp. 111–144.
- Belfort, M. and Perlman, P.S. (1995) *J. Biol. Chem.*, **270**, 30237–30240.
- Goodrich-Blair H. and Shub, D.A. (1996) *Cell*, **84**, 211–221.
- Marshall, P. and Lemieux, C. (1992) *Nucleic Acids Res.*, **20**, 6401–6407.
- Perrin, A., Buckle, M. and Dujon, B. (1993) *EMBO J.*, **12**, 2939–2947.
- Côté, V., Mercier, J.-P., Lemieux, C. and Turmel, M. (1993) *Gene*, **129**, 69–76.
- Schapira, M., Desdouets, C., Jacq, C. and Perea, J. (1993) *Nucleic Acids Res.*, **16**, 3683–3689.
- Turmel, M., Mercier, J.-P., Côté, V., Otis, C. and Lemieux, C. (1995) *Nucleic Acids Res.*, **13**, 2519–2525.
- Bryk, M., Belisle, M., Mueller, J.E. and Belfort, M. (1995) *J. Mol. Biol.*, **247**, 197–210.
- Loizos, N., Silva, G.H. and Belfort, M. (1996) *J. Mol. Biol.*, **255**, 412–424.
- Turmel, M., Boulanger, J., Schnare, M.N., Gray, M.W. and Lemieux, C. (1991) *J. Mol. Biol.*, **218**, 293–311.
- Gauthier, A., Turmel, M. and Lemieux, C. (1991) *Curr. Genet.*, **19**, 43–47.
- Turmel, M., Côté, V., Otis, C., Mercier, J.-P., Gray, M.W., Lonergan, K.B. and Lemieux, C. (1995) *Mol. Biol. Evol.*, **12**, 533–545.
- Marshall, P. and Lemieux, C. (1991) *Gene*, **104**, 241–245.
- Liu, S.-L., Hessel, A. and Sanderson, K.E. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 6874–6878.
- Toda, T. and Itaya, M. (1995) *Microbiol.*, **8**, 1937–1945.
- Liu, S.-L. and Sanderson, K.E. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 1018–1022.
- Honeycutt, R.J., McClelland, M. and Sbral, B.W.S. (1993) *J. Bacteriol.*, **175**, 6945–6952.
- Katayama, S., Dupuy, B., Daube, G., China, B. and Cole, S.T. (1996) *Mol. Gen. Genet.*, **251**, 720–726.
- Kaneko, T., Matsubayashi, T., Sugita, M. and Sugiura, M. (1996) *Plant Mol. Biol.*, **31**, 193–201.
- Schmidt, K.D., Tummeler, B. and Romling, U. (1996) *J. Bacteriol.*, **178**, 85–93.
- McCracken, D.A., Nadakavukaren, M.J. and Cain, J.R. (1980) *New Phytol.*, **86**, 39–44.
- Keller, M.D., Selvin, R.C., Claus W. and Guillard R.R.L. (1987) *J. Phycol.*, **23**, 633–638.
- Buccheim, M.A., Lemieux, C., Otis, C., Gutell, R.L., Chapman, R.L. and Turmel, M. (1996) *Phyl. Mol. Evol.*, **5**, 391–402.
- Turmel, M., Gutell, R.R., Mercier, J.-P., Otis, C. and Lemieux, C. (1993) *J. Mol. Biol.*, **232**, 446–467.
- Genetics Computer Group (1994) Sequence analysis software package, Version 8. Genetics Computer Group, Madison, WI.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
- Livingstone, C.D. and Barton, G.J. (1993) *Comp. Appl. Biosci.*, **9**, 745–756.
- Barton, G.J. (1993) *Protein Engng*, **6**, 37–40.
- Olsen, G.J., Matsuda, H., Hagstrom, R. and Overbeek, R. (1993) *Comp. Appl. Biosci.*, **10**, 41–48.
- Dujon, B., Belfort, M., Butow, R.A., Jacq, C., Lemieux, C., Perlman, P.S. and Vogt, V.M. (1989) *Gene*, **82**, 115–118.
- Dürrenberger, F. and Rochaix, J.-D. (1991) *EMBO J.*, **10**, 3495–3501.
- Lonergan, K.M. and Gray, M.W. (1994) *J. Mol. Biol.*, **239**, 476–499.
- Michel, F. and Westhof, E. (1990) *J. Mol. Biol.*, **216**, 585–610.
- Leung, D.W., Chen, E. and Goeddel, D.V. (1989) *Technique*, **1**, 11–15.
- Gimble, F.S. and Thorner, J. (1993) *J. Biol. Chem.*, **268**, 21844–21853.
- Gimble, F.S. and Stephens, B.W. (1995) *J. Biol. Chem.*, **270**, 5849–5856.
- Aggarwal, A.K. (1995) *Curr. Opin. Struct. Biol.*, **5**, 11–19.
- Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I. and Aggarwal, A.K. (1995) *Science*, **269**, 656–663.
- Marshall, P., Davis, T.B. and Lemieux, C. (1994) *Eur. J. Biochem.*, **220**, 855–859.
- Henke, R.M., Butow, R.A. and Perlman, P.S. (1995) *EMBO J.*, **14**, 5094–5099.
- Côté, M.-J. and Turmel, M. (1995) *Curr. Genet.*, **27**, 177–183.